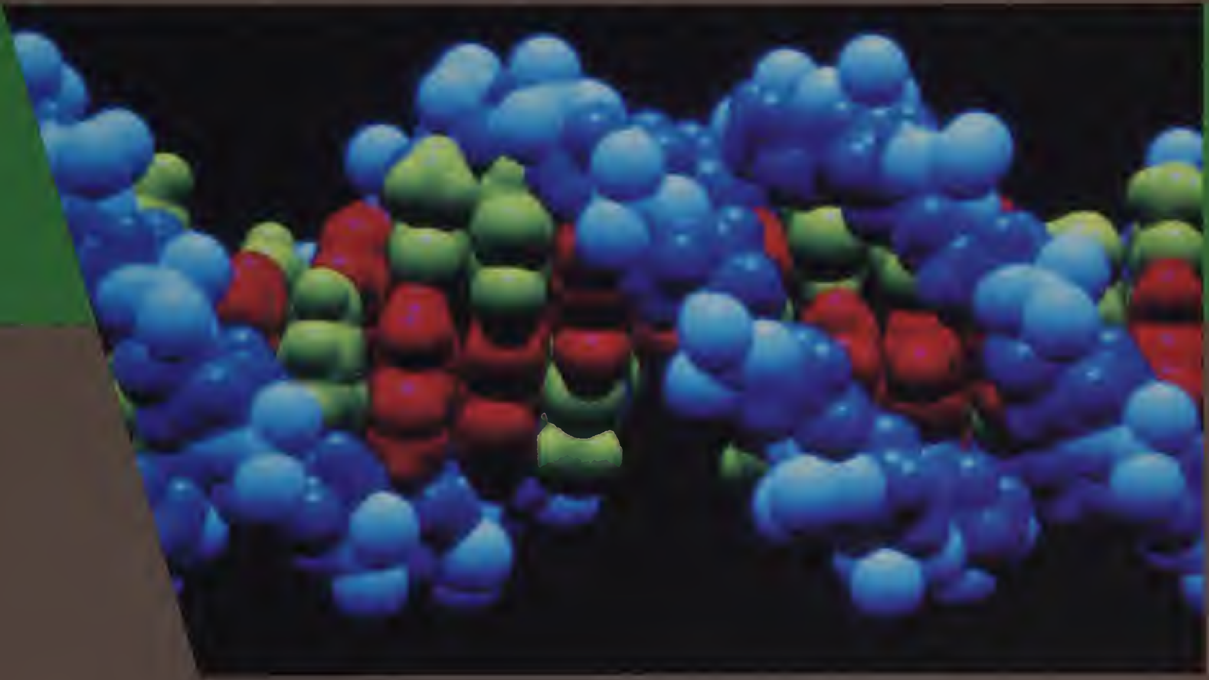
Report of Panel 3

Long Range Plan

National Library of Medicine

Obtaining Factual Information from Data Bases



U.S. Department of Health and Human Services
Public Health Service
National Institutes of Health

Report of Panel 3

Long Range Plan

National Library of Medicine

Obtaining Factual Information from Data Bases

U.S. Department of Health and Human Services
Public Health Service
National Institutes of Health

December 1986

Members and Staff of Panel 3 Obtaining Factual Information from Data Bases

Chairperson
Ruth Davis, Ph.D.
President
Pymatuning Group, Inc.
Arlington, Virginia

Members
Rachael Anderson, M.S.
Health Sciences Librarian
Columbia University
New York, New York

David H. Brandin
President
Strategic Technologies, Inc.
Los Altos Hills, California

James Burrows
Institute for Computer Science and Technology
National Bureau of Standards
Gaithersburg, Maryland

Robert Lee Chartrand, M.A.
Senior Specialist in Information
Policy and Technology
Congressional Research Service
Library of Congress
Washington, D.C.

Peter Friedland, Ph.D.
Senior Research Associate
Knowledge Systems Laboratory
Stanford University
Palo Alto, California

Robert E. Kahn, Ph.D.
Consultant
Information Processing Techniques Office
Department of Defense
Advanced Research Projects Agency
Arlington, Virginia

Joshua Lederberg, Ph.D.
President
Rockefeller University
New York, New York

Robert U. Massey, M.D.
Dean
University of Connecticut
School of Medicine
Farmington, Connecticut

Daniel R. Masys, M.D.
Chief
International Cancer Research Data Bank
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Allan M. Maxam Ph.D.
Assistant Professor of Biological Chemistry
Harvard Medical School
Dana-Farber Cancer Institute
Boston, Massachusetts

Gerard Piel, D.Sc.
Chairman of the Board
Scientific American
New York, New York

Richard J. Roberts, Ph.D.
Senior Scientist
Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Elmer V. Smith
Director
Canada Institute for Scientific
and Technical Information (CISTI)
National Research Council
Ottawa, Canada

Willis Ware, Ph.D.
Corporate Research Staff
The Rand Corporation
Santa Monica, California

Ronald L. Wigington, Ph.D.
Director
Chemical Abstracts Service
Washington, D.C.

NLM Staff
Sean P. Donohue, M.P.A.
Executive Secretary

Henry Kissman, Ph.D.
Resource Person

Consultants to Panel 3

Col. Andrew A. Aines
(U.S. Army, Retired)
Springfield, Virginia

William D. Carey
Executive Officer
American Association for the Advancement of Science
Washington, D.C.

Paul R. DeRensis, J.D.
Chairman
Section on Tort Liability for Use of Computer Systems
American Bar Association
Boston, Massachusetts

Vincent F. Guinee, M.D.
Chairman
Department of Patient Studies
Coordinator
International Cancer Patient Data Exchange System
M.D. Anderson Hospital and Tumor Institute
Houston, Texas

Warren J. Haas
President
Council on Library Resources
Washington, D.C.

Lawrence G. Hunsicker, M.D.
Associate Professor
Department of Internal Medicine
University of Iowa
Iowa City, Iowa

Laurence H. Kedes, M.D.
Professor of Medicine
Stanford University School of Medicine
Palo Alto, California

Donald W. King, M.S.
President
King Research, Inc.
Rockville, Maryland

Robert B. Lanman, J.D.
Office of the General Counsel, DHHS
NIH Legal Advisor
National Institutes of Health
Bethesda, Maryland

Miranda Lee Pao, Ph.D.
Associate Professor
Matthew A. Baxter School of Information and Library Science
Case Western Reserve University
Cleveland, Ohio

Robert D. Poling, J.D.
Specialist in American Public Law
Library of Congress
Washington, D.C.

Contents

Report of Panel 3

Long Range Plan

1 Background and Context 6

2 NLM Programs and Recent Accomplishments 8

The Hepatitis Knowledge Base 8
Online Reference Works 8
Factual Data Bases for Toxicology 9

3 A Vision of the Future 12

4 Major Issues and Future Directions 13

Medical Practice-Linked Data Bases 13
Patient-Record Data Bases 15
Biomedical Research-Oriented Data Base 17
Expert and Modeling Systems 20
Technologies in Support of Factual Data Bases 24
Societal and Institutional Considerations 27

5 Observations and Recommendations 30

Observations 30
Recommendations 30

References 34

Appendix A
Factual Data Bases in Basic Research 35
Overview
Uses and Users
Data Base Maintenance
Limitations
The Future

Appendix B
Emerging Technologies:
A View to the Future of Factual Data Bases 38

Appendix C
NLM Planning Process 40

Background and Context

The charter of the NLM (National Library of Medicine) affords considerable latitude in the information and knowledge-based services that the Library may provide to the medical research and health-care communities. That latitude, however, imposes responsibility for determining, in conjunction with existing and potential users, which services are needed most and how best to provide them. This report explores the current and future roles of the Library in archiving and transferring current biomedical knowledge through factual data bases.

With the advent of increasingly sophisticated electronic information processing tools, the traditional concept of data base management has expanded from the manipulation of raw numbers to include a wide range of knowledge-based capabilities. A knowledge base is defined as a comprehensive body of information on a given subject. It comprises a variety of factual information in addition to bibliographic citations and provides a current consensus of experts on the subject. The information in a knowledge base may take several forms, including expository (declarative statements), procedural (rules for decision making), and inferential (derived from reasoning).[1] To be useful, a computerized knowledge base must be immediately responsive to inquiries and should afford access to different levels of data.

In the context of this report, then, factual data bases represent structured knowledge that is acquired, stored, processed, and disseminated using automated electronic systems. Such data bases differ from bibliographic retrieval systems in their capacity as fact providers rather than fact locators. The differences between factual data bases and bibliographic systems are often substantial, including the methods by which the two are constructed, the safeguards needed in choosing their content, and the need for rigorous assessment of their quality and timeliness.

For the purposes of this discussion, therefore, factual data bases have been taken to encompass all collections of data, signals, and information other than bibliographic records that satisfy a narrow definition of a factual data base, such as a collection of measurements of physical or chemical properties of a biologic system. Also included are those that fill a wider perspective by including observations, ideas, and opinions. Thus, electronic representations of graphic images or audio recordings are considered factual data bases, as is the online dissemination of an editorial statement that represents the consensus of experts in an area of biomedicine.

The Panel recognizes the lack of clear boundaries between full-text bibliographic retrieval systems and factual data bases. The electronic transfer of textual information from a factual data base is electronic publishing, and presents the same issues of acquisition, content review, and dissemination as traditional paper-based publication. Two common properties of the diverse group of information sources known as factual data bases keep them within the Panel's purview: First, all have been acquired, stored, and distributed solely or principally by electronic means. Second, the data themselves hold the information content, rather than acting as citations or pointers to other sources.

It should be noted that this report excludes from consideration those types of data bases treated directly by other panels: biomedical and scientific bibliographic data bases, with their accompanying citations and abstracts, and educational or instructional factual data bases.

2

NLM Programs and Recent Accomplishments

The Hepatitis Knowledge Base

The development and use of electronic factual data bases on a broad scale is a new and rapidly evolving field, one with significant potential for application at NLM. From 1976 to 1981, the Lister Hill National Center for Biomedical Communications of the Library explored the creation of knowledge bases as prototype information-transfer systems for use by health practitioners.

The Hepatitis Knowledge Base was an outgrowth of that work. Originally assembled from a textbook chapter and 40 recent review articles,[2] its contents were augmented by a bibliometric process that selected high-quality articles from relevant journals. The question of obsolescence was circumvented by having experts analyze the selected articles, extract only new and important information, and revise the data base at frequent intervals. For ease of information retrieval, free-text search capabilities were incorporated.

Group consensus was regularly sought on the contents of the Hepatitis Knowledge Base. A 10-member panel of nationally prominent experts in the field of viral hepatitis reviewed the original draft document and all subsequent editions. Newly available computer conferencing techniques were used to link the experts with each other and with Library staff. In this way, proposed changes were transmitted, discussed, and agreed on without costly and time-consuming face-to-face meetings.

Paralleling research and development was a formative evaluation. The methods used to construct the data base and maintain its quality were assessed. In addition, a year-long field test examined the performance of the online access system at a variety of sites.

The evaluation demonstrated that it is feasible to construct a high-quality, full-text data base. The approach to data reduction and quality control was, in fact, effective and efficient. Further, computer conferencing to obtain group consensus proved to be a practical way of maintaining the currency and accuracy of the data base content. Finally, an evaluation of the online access methodology showed that, for 85 to 95 percent of user queries, the Hepatitis Knowledge Base was able to locate the information sought within the first few paragraphs displayed.

Online Reference Works

The Lister Hill Online Reference Works program was established to evaluate strategies for the automated retrieval of textual information from existing medical literature. In particular, the program seeks to provide low-cost, efficient information retrieval systems that do not require specially structured knowledge bases.

An allied research area is the design of an electronic writing environment for preparing and revising reference works and scholarly texts. With the introduction of high-density storage devices and powerful hardware, it has become possible to envision a ‘‘scholar’s workstation’’ that could serve both the student and the author as an integrated information resource.

Toward those ends, Lister Hill is currently collaborating on a project with the Welch Medical Library and the Johns Hopkins School of Medicine. The project involves developing an online reference work from an 1,800-page, 11-million character textbook on inherited disorders. Called *Mendelian Inheritance in Man* (MIM),[3] the publication is now in its sixth edition and is continually updated by the author.

The project's immediate objectives are to establish both visual (a human gene map) and text word means of information retrieval. Techniques previously developed by Lister Hill for IRX (Information Retrieval Experiment) are being integrated into OMIM (Online Mendelian Inheritance in Man) for user evaluation in clinical and library settings.

A set of tools for authors developed for online text preparation will be used to produce the next edition of the reference. The linkage of related data bases has begun with work to incorporate the human gene map and a set of clinical synopses in the text.

For the future, a "scholar's workstation" is planned, and will include optical disk storage, high-resolution video display, and computer network access. Artificial intelligence tools—such as syntactic parsers (programs that decompose phrases and sentences into logically related word groups), natural language processors, and frame-based indexing methods—will be added in later stages of the project.

Factual Data Bases for Toxicology

The NLM Division of Specialized Information Systems (SIS) is responsible for the TIP (Toxicology Information Program). TIP was established in 1967, in response to the 1966 President's Science Advisory Committee report on *The Handling of Toxicological Information.*[4] The Committee pointed to "an urgent need for a coordinated and more complete computer-based file of toxicological information than any currently available, and further, that access to this file should be more generally available to all those legitimately needing such information." TIP was created to meet the need and continues to develop innovative ways of providing toxicology information to its growing user community.

The program's objectives are to (1) create and maintain automated toxicology data banks and (2) disseminate toxicology information through a number of services, including publications, individual query response, and online information retrieval. TIP's early efforts were limited to publications and responding to queries. During that time, rapidly changing computer and telecommunications technologies were investigated for potential application to automated toxicology information systems.

Interactive online retrieval services now play the major role in TIP's activities. Following the pioneering development of MEDLINE by the Library, the earliest large-scale online bibliographic retrieval system, TIP unveiled TOXLINE, the first online retrieval service for toxicology literature, in 1972. The same year saw the development of CHEMLINE, an online factual data base for chemical nomenclature. CHEMLINE's purpose is to facilitate the searching of TOXLINE and other information sources.

In the late 1970's, TIP made publicly available two online factual data bases for toxicology: TDB (Toxicology Data Bank) and RTECS (the Registry of Toxic Effects of Chemical Substances). TIP builds and maintains TDB. RTECS is produced by the National Institute for Occupational Safety and Health as a publication. It is also maintained as an online service through the NLM MEDLARS (Medical Literature Analysis and Retrieval System) network. The most recent TIP online factual data bases are DIRLINE (Directory of Information Resources Online) and HSDB (Hazardous Substances Data Bank).

The remainder of this section describes these TIP-developed factual data bases in more detail.

CHEMLINE is an online chemical dictionary and directory file. It allows users to identify a chemical substance, determine which NLM files contain related information, and formulate a search strategy for those files. Currently, CHEMLINE contains over 650,000 records of chemical substances. It is updated bimonthly and regenerated at least once a year.

Recently, 14,800 drug names taken from the *United States Adopted Names and the United States Pharmacopiea Dictionary of Drug Names* have been added to over 5,500 CHEMLINE records. The addition of ring structure information to records for cyclic compounds in CHEMLINE continues with 9,000 records having been enhanced to date.

RTECS is an online factual data base built and maintained from data provided by the National Institute for Occupational Safety and Health. Recently, the file was enriched by adding Chemical Abstracts Service Registry Numbers to RTECS records that did not have them. These identification numbers are crucial for unambiguous data retrieval and for matching RTECS records with those in other files. Some 4,700 records have been enhanced in this way; another 14,000 remain to be processed. RTECS now contains over 76,000 records.

TDB is another online factual data base describing chemical substances that may be hazardous and may have significant human exposure potential. TDB records include information on pharmacology and toxicology, manufacturing and use, environmental and occupational exposure, and chemical and physical properties. Data are extracted from tertiary sources such as monographs and handbooks, as well as from primary literature. Completed records are evaluated by the Scientific Review Panel, which is composed of toxicologists, environmental scientists, and industrial hygienists. The online file currently contains 4,100 records, each consisting of 96 data elements.

HSDB, the newest of TIP's factual data bases, describes the same 4,100 records contained in TDB. However, HSDB expands on TDB by providing information on environmental fate and exposure, standards and regulations, monitoring and analysis, and safety and handling. Data are extracted from TDB source materials and various other sources, such as government documents and material safety data sheets. HSDB is also reviewed by the Scientific Review Panel.

DIRLINE refers MEDLARS users to organizations and other sources that can provide information in specific subject areas. DIRLINE has been available online since August 1984. At present, DIRLINE receives records from the Library of Congress' NRC (National Referral Center) data base and from the Department of Health and Human Services' NHIC (National Health Information Clearinghouse) data base. The NRC component contains some 14,000 records, while the NHIC file has approximately 950 records. New file segments derived from a Food and Drug Administration list of poison control centers and from a National Institute for Alcohol and Drug Abuse directory of regional drug and alcohol centers are being prepared for addition to DIRLINE in fiscal 1987.

Recent TIP developments include the creation of TOXNET (Toxicology Data Network), an integrated software system that facilitates the building and searching of factual data bases. Micro-CSIN, a newly developed microcomputer-based Chemical Substances Information Network, speeds retrieval of information from disparate data bases residing in various online systems.

3 A Vision of the Future

The past two decades have seen an amazing proliferation of information and data processing technology. Over the years, NLM has rightly assumed a lead position in innovating and applying electronic factual data bases to an ever-growing pool of biomedical knowledge. Looking forward to the 21st century, the director of the Library offers a vision of the future for biomedical information transfer in *The Distant Goal.*

- All health professionals in the United States will be able to obtain computer-based, practice-linked automated information assistance.
- NLM will provide access to computer-based expert consultant systems and raw appendiceal files (when they exist).
- Patient records will be stored in a machine-readable form for a variety of purposes.
- A substantial cadre of well-trained information specialists will be employed in schools, libraries, industries, and hospitals.
- Virtually all. . .biomedical specialists will have [compatable computers] at home and at work. [Those] machines will have access to public and local. . .networks,. . .protocol and computer language compatibility. . .will be provided cheaply.
- Personal computers and communications networks will provide access to thousands of online data bases, information bases, knowledge bases, and expert consultant services.
- Most professionals will perform significant amounts of work-time activities at home. . .from 10 to 85 percent of the professional work week.
- Interlinked systems. . .with individual patient care data,. . .literature-based consulting systems, and continuing education systems will be utilized by the health-care professional.
- Records of individual patient history, treatment, and observations will be available electronically. . .access. . .[will be] permitted only with the active agreement of the patient.
- The best [research] articles [will]. . .include machine-readable appendices that provide the raw data in [a] reference library information center.
- The ability for free-text/full-text retrieval will be universal, but special libraries will have additional automated expert search systems.
- Record centers, including some biomedical libraries, will hold. . .patient-care records. . .[that] include radiant images, physiological data, photographs,. . .interrogative patient history, physical examination, treatment, and interval medical notes. Such services. . .[will also be available] to others at remote sites via fee-for-service dial-in access arrangements.

It was in the context of these insights that the Panel deliberated the major issues and future directions in obtaining factual information from data bases.

Major Issues and Future Directions

4

Medical Practice-Linked Data Bases

Computer-based factual data bases are new tools; their preparation and maintenance do not replace the traditional mandate of the Library. Neither the explosion of medical information nor the technology to deal with it existed in the past. Therefore, the surge forward in this area presents the Library and the medical profession with an opportunity that will largely require obtaining new resources rather than reprogramming existing ones.

Although the Library may expect to build and maintain some factual data bases, more of its resources in the future are likely to be devoted to distributing biomedical information whose content is the responsibility of other organizations. PDQ (Physician Data Query) highlights many of the pertinent issues that the Library faces in developing and distributing data bases linked to medical practice.

PDQ consists of three interlinked files of cancer-related information that first became available as an online service in 1984.[5] PDQ's information content is assembled and maintained by the NCI (National Cancer Institute). NCI bears full responsibility for the content of this data base, which provides recommendations that can directly affect patient care for life-threatening diseases.

NLM functions as a central online distribution facility for PDQ's wide and previously established user community. Through an intra-agency agreement, NCI refunds the Library the costs of making PDQ content available through MEDLARS, providing a multiuser data base management system for online retrievals.

The costs of future factual data bases will likely be borne jointly by the Library and other institutions, public and private, that wish to benefit from the Library's substantial distribution network and skill at building data bases. Such joint ventures, when consistent with the Library's legislative charter, should be encouraged. In fact, they should be planned so the number and types of factual data bases available to the users of the Library can be increased without consuming an ever-greater share of the Library's budget.

NCI developed the file structures and retrieval software for PDQ. The result is a functional stand-alone data base. An extensive set of cancer-specific index terms and synonyms are provided in PDQ. These terms, however, are not linked to other data bases or vocabularies (i.e., MeSH (Medical Subject Headings)).

Similar efforts that other NIH institutes undertake in the future may, if developed independently, also fail to provide the consistency necessary to accommodate queries across data bases. The Library is the most suitable organization to provide guidance and standards for the design of such systems.

PDQ's user-friendly design was originally intended for health professionals with little or no experience in using computerized information retrieval systems. The system displayed a list of options, called a menu, on the computer screen, and the user selected one. Systems like this work well for the occasional user. However, they tend to frustrate the more experienced user accustomed to entering command statements to locate desired information. Consequently, an expert-user mode is gradually being developed and installed in PDQ.

Ideally, future factual data bases should be flexibly designed to satisfy a variety of user needs. In particular, future systems would benefit from the ability to display the equivalent search statement in command language, menu choices, and graphic depictions of the logic contained in the user request.

Compared with a potential audience of over 200,000 physicians nationwide, PDQ's 350 users per month is relatively small. NCI's preliminary evaluation of this problem points to two major impediments.

First, there is a relatively low level of awareness in the medical community that the system exists. In 1985, NCI surveyed communities where it has Community Clinical Oncology Programs. The results showed that approximately 80 percent of the cancer specialists in those communities knew about PDQ, but only 30 percent of the other physicians surveyed had heard about the system.

Second, the information-seeking habits of physicians do not, in most cases, include the routine use of computerized practice assistance. PDQ contains information that cannot be easily accessed anywhere else, and it is targeted to both specialists and nonspecialists. Yet, of the cancer specialists aware of PDQ, only 1 in 10 had actually used the system or had a search performed for them.

It is clear that, if the routine use of computerized information to improve patterns of medical care is to become a reality, the health professions must be educated in modern information retrieval methods. As a provider of factual data base services, the Library should bear a major responsibility for fostering such efforts.

For physicians whose practice patterns and information-seeking habits are already set, the economics of attempting to change their behavior through advertising and promotion are not cost-effective. The opportunity for change will come, instead, as more students and young professionals enter the field having gained some experience with computers during their training.

Until then, a more productive avenue for the Library may be to promote the use of factual data bases by teaching hospitals and standard-setting organizations, such as the Veterans Administration and large corporate health-care organizations. The potential of such organizations to influence the behavior of professionals supplies a powerful complement to educational efforts in health science curricula.

It is possible, however, that the transition to widespread use of computers in routine medical practice will occur very suddenly, driven by legal precedents of liability accruing to practitioners who do not use the best, most up-to-date sources of information. A similar phenomenon occurred within several years of the introduction of computerized tomography in the evaluation of central nervous system disorders.

PDQ is licensed to two domestic commercial vendors and one foreign vendor. Although the system's information content is not subject to copyright, NCI uses the license agreement as a quality assurance mechanism, releasing updated tapes only if the vendor has satisfied certain criteria, such as accurate and timely maintenance of the data base. The license agreements also help defray the costs of NCI's technical information services.

The Library's factual data base services will exist in an increasingly heterogeneous environment of distribution systems. There will be many variations of online, centrally accessible information, as well as subsets of information provided on different media. For example, NCI has received proposals to incorporate PDQ information into optical disk retrieval systems.

To the extent practical within its public mandate, the Library should develop licensing arrangements with other domestic and foreign organizations, public and private, to promote the dual purposes of wider dissemination and cost reimbursement. Such arrangements, however, should be contingent on the Library's retention of the right to assess the quality of the information delivered by the licensee. Whenever feasible, the costs of quality assurance should be borne by the licensee.

Patient-Record Data Bases

The clinical importance of detailed information about individual patients is generally unchallenged. Most physicians rely on a patient history in making a diagnosis, devising a treatment plan, and recording therapeutic outcomes. The patient record, however, is used quite differently from the way medical practice-linked data bases are. As a result, the two sources of information must be sharply differentiated.

The automated handling of patient records poses many problems. One is confidentiality of personal data; another is comparability and accuracy of data entries from different sources. Policy and management decisions must reside primarily with health-care providers, not with the medical research community. Using patient records in a research context is a separate issue, but one the Library can deal with, once it has addressed the problems of protecting privacy and ensuring accuracy.

The term ''patient record'' implies several different uses in the medical community. The three most common are considered here.

First, a patient's record is a history of medical encounters (including hospitalization) kept by health-care providers to document what is known about the patient's health, diagnoses, treatments, clinical courses, and data (including images and test findings) that support or negate any of these elements. Ordinarily, patient records are likely to be episodic, in the sense that each hospital or practitioner initiates a unique record rather than transferring one from a previous provider; tends to use language and iconic representations that are a vernacular rather than a standard vocabulary; does not observe uniform standards regarding completeness or relevance of entries; and may use terms that appear to be the same but that in fact have different meanings to those using them. Other sorts of nonstandardization (such as format) may also occur.

The result is that hospital charts and office records are of limited scientific value for the epidemiological study of illness and treatment. Moreover, they often have limited clinical value to both the practitioner and the patient.

Individual patient records necessarily contain a good deal of identifying information about the person that goes beyond what might be needed for scientific research. This raises the issue of confidentiality and patient privacy.

Although they do not currently exist, methods have been proposed for overcoming some of these obstacles and increasing at least the clinical usefulness of the patient record. One proposal, now under development, is a credit-card-sized, electronically written, and machine-readable record of medical encounters. The card would remain in the patient's possession, but could be used by a succession of providers, assuming uniformity in encoding and decoding devices.[6] Presumably, this technology would also require uniform or readily translatable language, format, and other standards for entry in the record. An alternative development is the centralized (and presumably standardized) collection and storage of individual medical records currently being organized by a number of commercial firms.

In addition, the gradual but persistent trend toward the corporate organization of health care will implicitly encourage uniformity in record keeping, if only for reasons of administrative efficiency. This trend will also complicate the process of protecting privacy and maintaining confidentiality, but may accelerate efforts to find technological solutions to those problems.

The second category of patient records comprises those contained in specialized data bases defined by disease entities or other research criteria. Now relatively common, such collections may be expected to assume greater importance in the future. Along with conventional tumor registries and groups of patient records such as the Duke cardiology data base, there are records of longitudinal surveys such as the Framingham study and randomized clinical trials whose value as data may well go beyond the purposes (and the investigators) originally intended for them. The opportunities for secondary analysis, historical comparisons at some later date, and the retrospective investigation of phenomena not known or suspected during the original data collection—all give rise to the recommendation that the Library should seriously consider playing an influential role in maintaining and archiving such data bases, as well as making them accessible to future investigators.

Privacy and confidentiality also remain thorny issues in this category of patient records. In longitudinal studies or studies comparing individuals across files, it is necessary to maintain unique identification, although not to retain information enabling someone to find an actual name, address, or other identifier. Technological advances in the next several years should do much to resolve such issues. Many other agencies and institutions face similar problems even more urgently then does the Library. Consequently, their solutions may be soon forthcoming. The Library may have opportunities to adopt or encourage the adoption of those solutions in the area of patient records.

Besides privacy and confidentiality, there are significant questions of standardization in collecting and encoding primary data, ordering and indexing what is collected, and above all, documenting the storage and retrieval programs so as not to impair the usefulness and validity of the data. Working with scientific and professional societies and government agencies involved in health care and research, the Library should develop standards for this category of data bases.

The third form of patient record consists of the clinical evidence supplied to substantiate a published analysis of some medical issue or phenomenon. Electronic publication of scientific research papers appears to be on the horizon. When available, this format will make it possible to include lengthy files of primary data that provide the evidence on which the investigator rests his case.

The scientific gains from electronic publication are obvious, and the problems accompanying them are somewhat less formidable. Privacy and confidentiality issues need not be serious unless there is some reason to link the published material with other types of information or other files on the same individual—an unlikely occurrence outside large studies. Questions of standardized nomenclature, record storage, and the like are no worse than for other types of patient records, and for many purposes, may be considerably less weighty.

Biomedical Research-Oriented Data Bases

Basic research in the life sciences is becoming increasingly dependent on automated tools to store and manipulate large amounts of data that describe the behavior of biologically important macromolecules. The ability to measure and change events occurring on a molecular level is particularly significant in the field of genetics, where the development of techniques to sequence, clone, and remodel DNA (Deoxyribonucleic Acid) is leading to control of life processes with a precision never before seen. Accordingly, the use of data bases in molecular genetics has become a field of compelling scientific promise.

A case study of the gene sequence data base called GenBank appears in Appendix A. It is only part, however, of a much larger and more complex array of information contained in data bases at many sites throughout the world.

The types of biogenetic information available parallel the hierarchy of biology itself: Levels of inquiry range from cells through successively smaller genetic units to base-pair sequences, as shown the adjacent figure. Although the list of information resources is intended to be representative rather than exhaustive, the figure does show that computerized collections of information exist at every level, each with its own distinct data base structure and unique indexing and retrieval methods.

Historically, each data base was developed to catalog, store, and support analysis of new genetic information at its own level. Pointers to information at other levels of biological structure are only now beginning to appear in the data bases.

The relatively isolated design of the various data bases contrasts sharply with the current research activity in molecular biology, where an investigator will commonly report findings involving data at the cellular, chromosomal, gene, amino acid, and DNA sequence levels within a single scientific paper. Additionally, the critical scientific questions being asked can often be answered only with the ability to relate one genetic level to another.

One critical contribution of computerized information to the advancement of biological knowledge is in the field of oncogenes (cancer-causing genes). Within the last few years, it has been discovered that genes found in malignant human cells can be transferred into normal cells and cause them to become malignant. The complete DNA sequence of many of these genes has been determined. By comparing these sequences with other known sequences present in data banks, it has been found that often they are very closely related, if not identical, to genes that are present in certain RNA tumor viruses that have been shown to induce cancer in mice and chickens.

Just three years ago, a highly significant breakthrough was made when it was discovered by computer analysis that one of these viral oncogenes called v-sis was almost identical in sequence to a normal human gene encoding a growth factor (platelet-derived growth factor). For the first time, a biochemical function could be postulated for a gene that was associated with the onset of cancer. This finding has led to a flurry of experimentation aimed at examining the precise mechanism by which this gene is able to induce cancer. Most important, it has shown that by looking for structural relationships between the DNA sequences of newly discovered genes and previously known genes, it is often possible to make enlightened predictions about the function of the new genes. Appropriate experiments can then be performed to test these predictions directly. Not only is this of enormous importance in the field of oncogenes, but it has greatly speeded research in many different biological arenas. The original breakthrough and the possibilities for future progress depend entirely on the existence of factual data bases and sophisticated computer programs for their analysis.

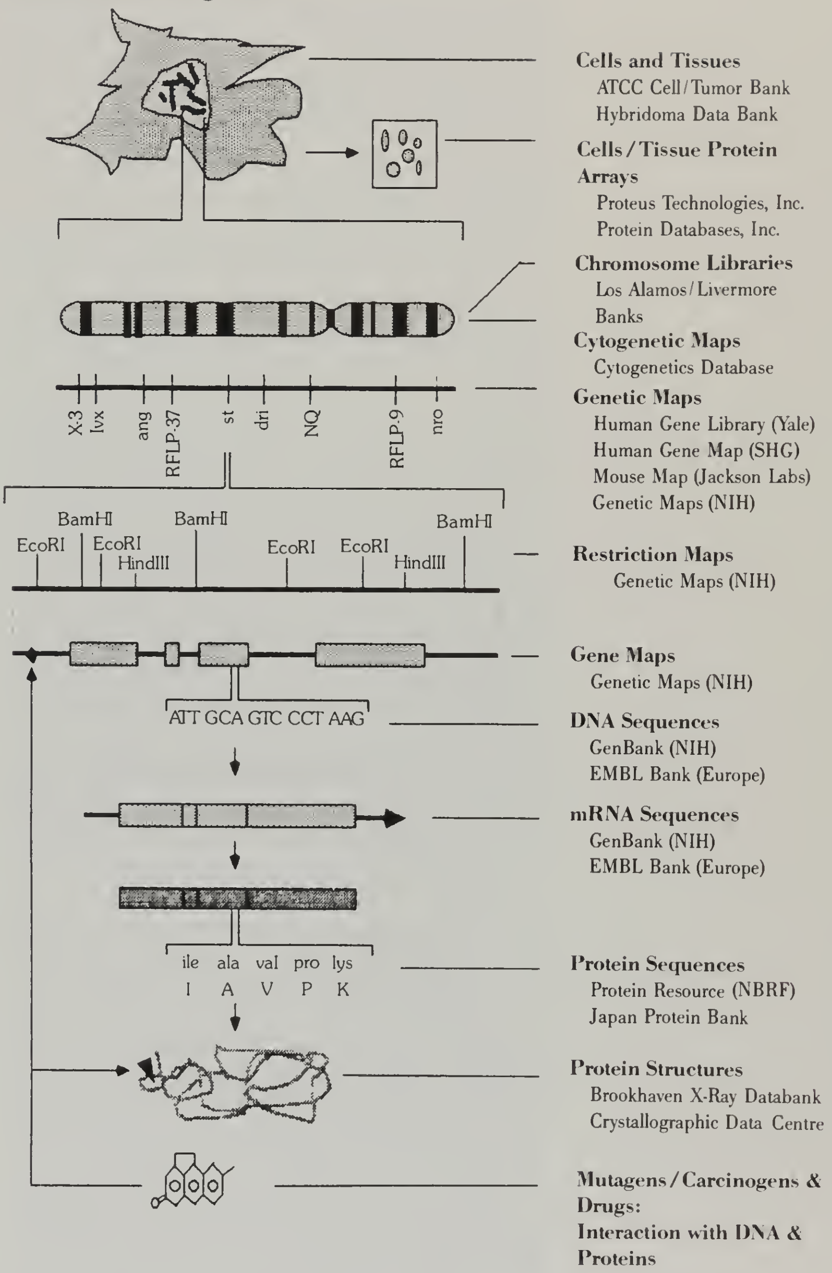
The general area of biogenetics is moving ahead rapidly. Serious proposals have been put forward to sequence the entire human genome and to map active chromosomal regions for each tissue type in different organ systems. Automated equipment to do this is being developed and will accelerate the acquisition of new data. Associated data bases already contain information essential to such work. Researchers are now sorting chromosomes using fluorescence techniques, building clone libraries of those chromosomes (collections of synthetically duplicated genetic material), and establishing relationships between structure and function.

Fundamental research in genetics permeates the life sciences. For instance, the prenatal diagnosis of blood disorders, such as sickle cell disease, has been made possible through newly acquired genetic knowledge. The production of therapeutic agents, such as interferons and interleukins, depends on DNA and protein sequence information assembled in accessible data bases.

The research-oriented information systems currently in place are adequate to ask low-level questions: Find the degree of similarity between base-pair sequences. The next questions are: What do the differences mean? Current data bases are being used to support modeling and theory, but the tools are very primitive, and no methods exist for automatically suggesting links across levels. There is a vacuum in the area of research into ways of using information by interconnecting various levels. The deeper understanding of biology will ultimately require making those connections.

Currently, no organization is taking the lead in promoting keys and standards by which the information from the related research data bases illustrated in the accompanying figure can be systematically interlinked or retrieved by investigators. Progress toward such interlinkages would be made in the short term with the development of conventions for uniform indexing and a thesaurus to translate identical and related terms.

Biology Knowledge Bases



Biomedical Data Bases in a Universal Hierarchy of Nature: cells—chromosomes—genes—proteins.

NLM currently plays a crucial role in this science. For example, much of the information about the rapidly expanding field of molecular biology is in the literature encompassed by MEDLINE. Although this literature is indexed by the Library's MeSH, scientists within the field could benefit from additional indexing methods.

A singular and immediate window of opportunity exists for the Library in the area of molecular biology information. Because of new automated laboratory methods, biological data are accumulating far faster than they can be assimilated into the scientific literature. The problems of scientific research in the field of molecular biology are increasingly problems of information science. The full potential of the rapidly expanding information base of molecular biology will be realized only if an organization with a public mandate such as that of the Library takes the lead to coordinate and link related research data bases and make them easily accessible to the U.S. and international research communities.

The long-range goal of identifying and retrieving related data and concepts will eventually require natural language indexing and powerful search capabilities. The complex ideas embodied in the research are not amenable to electronic retrieval by the search technology now available. Unfortunately, workable natural language capabilities are at least 10 years distant, and require substantial improvements in both software function and hardware speed.

Expert and Modeling Systems

Another approach to automating information retrieval at different levels of integration is to define and model the processes used by librarians, information specialists, and other expert searchers who currently perform query analysis and source selection. The Library would be uniquely suited to supporting, in two parts, this movement toward the future of data base management.

The first part would be funding the development of an expert system that models the methods employed by an excellent medical reference librarian. The second part would consist of research into ''intelligent introspection'' systems that can, alone or in partnership with human scientists, examine medical and scientific data bases to form correlations and linkages.

An important use of factual data bases is to provide the basis for expert systems such as those developed experimentally for making diagnoses and determining treatment regimens in medicine. The idea is well illustrated by the MYCIN and ONCOCIN programs developed at Stanford University.[7]

MYCIN, an antibiotic-selection assistance program, demonstrated that the way physicians diagnose disease and prescribe therapy could be effectively modeled on a computer. However, it was never placed in a hospital setting because of perceived economic and behavioral problems that would have limited its utility and acceptance.

First, physicians did not generally see the need for the system; they believed that their approach to antibiotic selection was satisfactory, or that the investment of time required did not justify the benefit received. Second, the physicians found the mechanics of computer terminal use annoying. As one expert put it, ‘‘If the screen flickers, the doctor won’t use it.’’

In response, the artificial intelligence group at Stanford chose a follow-on task considered more likely to succeed: decision assistance with experimental cancer treatment protocols. Oncology protocols are complex, and few physicians see themselves as experts on any given protocol. Even those who are experts have voiced a need for assistance in conducting clinical trials.

The ONCOCIN model is that of a cancer expert who provides advice and explains the basis for that advice. ONCOCIN’s goals are to (1) give excellent advice about patient management, (2) improve the ease and accuracy of data management, (3) avoid the use of an intermediary between physician and computer, and (4) make the system suitable for dissemination to an office practice environment.

ONCOCIN’s decision-support component is largely based on IF-THEN rules, but also invokes several other types of reasoning. The original concept has evolved into a set of research projects: (1) ONCOCIN itself, which gives advice while collecting and storing clinical data; (2) OPAL, a knowledge-acquisition system designed to help the expert describe his protocol to ONCOCIN; and (3) ONYX, a system intended to deal with the 5 to 10 percent of the cases where the existing set of protocol rules is insufficient to support a decision. In those cases, the system attempts to make a causal model of events and relate them to basic medicine and biology. The modeling in ONYX simulates the normal responses of human organ systems to toxic stimuli.

Formal evaluation of ONCOCIN’s advice has shown that it performs as well as clinicians caring for cancer patients in a busy university oncology clinic. It will soon undergo testing on a small scale in an office-based practice.

ONCOCIN is a good example of a second-generation system. It resolved the technical and user acceptability problems of earlier systems. The computer hardware necessary for its use has just now become sufficiently economical to allow more widespread dissemination. Later generations will incorporate more advanced hardware and software. A major improvement will be the integration of that hardware with video disk devices, which will allow visual images to be displayed. The software is currently appropriate for handling most cases. However, that needed for modeling based on deeper representations of knowledge is perhaps 5 to 10 years away.

Because the technology will certainly be cost-effective and manageable enough for systems such as ONCOCIN to be used by all physicians in the foreseeable future, the impediments to adoption are mainly behavioral and social. Suppose a computer-based system makes better judgments than the average physician? Do human skills atrophy when physicians depend entirely on the system? Also, untested legal issues will have to be faced: If ONCOCIN is recognized as a standard, does a physician incur legal liability for care delivered without consulting the system? These questions must be explored for ONCOCIN and all medical expert systems.

NLM has a continuing commitment to expert system development through inhouse efforts, grants, and other support. This includes systems for use in medical education, chemical spill decision making, and the Library's artificial intelligence projects in rheumatology and blood coagulation, known as AI-RHEUM and AI-COAG respectively.[8]

Decision-support systems in areas other than medicine have enjoyed commercial success. Most current examples exist in the financial services area. Medical systems have come only from academic research environments. However, once their usability has been proven, the commercial potential of medical expert systems appears extensive.

Consequently, the utility of medical expert systems needs to be formally demonstrated. The agency that sponsors the research and development of such a system should also underwrite the important but potentially more expensive step of conducting a field test large enough to evaluate the system's worth in actual practice.

When medical expert systems become available to support decision making about standard treatment programs, the Library could play an important role by housing standard therapy guidelines in electronic format, just as it now archives standard printed textbooks. The Library should also take the lead in accumulating and constructing the factual data bases required to support these systems.

Some Library factual data bases will serve as primary sources of knowledge for expert systems designed by others. An example is a planned expert system for emergency responses to the spill of hazardous chemicals, which will use the Library's Hazardous Substances Data Bank as one of its principal information sources.

Computer modeling of life processes is similarly linked to and dependent on factual data bases. Both quantitative (standard numerical) and qualitative (causal, artificial-intelligence based) modeling systems have important roles to play in testing and expanding the utility of factual data bases.

It is becoming widely recognized that computer-based modeling activities that attempt to predict biological activities of chemicals based on known activities of structurally related chemicals can play important roles in developing data for such varied activities as risk assessment, synthesis of new pharmaceutical or agricultural products, and the reduction in the number of animals needed for biological research and testing. The latter product of computerized structure-activity modeling is of great importance to the continuing effort by NIH—and our society in general—to decrease death and suffering of laboratory animals.

The Library should, therefore, foster the development and operation of such modeling systems through suitable organization of the content and structure of its data bases; facilitate data transfer to modeling systems; and, where relevant to the Library's mission, study the development of better and more reliable modeling systems, particularly through the application of artificial-intelligence methodology.

Once the tools of medical expert system building have been developed, tested, and validated, the commercial viability of some systems would drive their continued development by private vendors. Likely examples include expert systems for clinical decision assistance regarding drug therapy for hypertension, diabetes, and the prevention of adverse drug interactions in patients with multiple medical problems. Development of these kinds of systems might be supported fully or in part by the pharmaceutical industry.

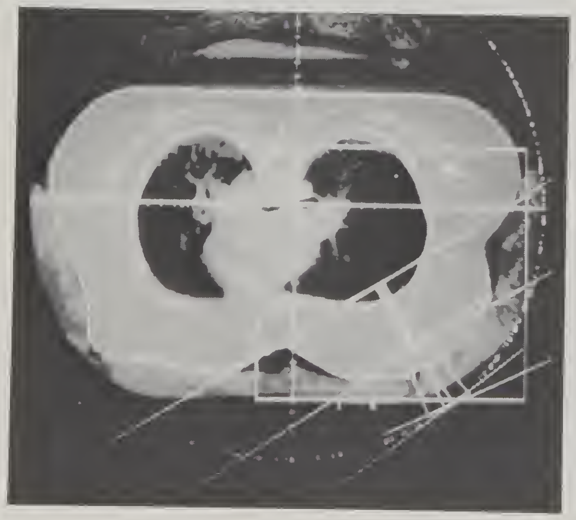
In other areas, the development of medical expert systems would represent a poor economic risk for private enterprise. In such cases, which are conceptually similar to Federal Government sponsorship of "orphan drugs," the Library should undertake system development for the public good. Examples here might include clinical decision assistance regarding management of rare but life-threatening diseases. The Library's artificial-intelligence based AI-COAG project, for instance, provides interactive, practical management assistance for treatment of proven or suspected bleeding disorders.

Regardless of the potential for commercial exploitation, however, it must be recognized that the technologies of medical expert-system design are currently under development in a few specialized centers, largely supported by the Library's extramural grants program. Continued Library support for this promising area is essential.

Technologies in Support of Factual Data Bases

The character of factual data bases and their utility will be altered by advances in computer hardware and software, communications technology, and networking. Each of these areas will present windows of opportunity as well as impediments to achieving Library goals. In addition, significant commercial market forces will assist or hamper progress toward ends in which the Library has a distinct interest.

Library policy is unlikely to affect the pace of commercial development. Nevertheless, the Library should stay well informed about technological progress in those areas relevant to biomedical information. Toward that end, "Appendix B" reviews the emerging technologies likely to have an impact on the way factual data bases are designed and used in the coming decade.



Nontextual Signal Storage and Retrieval

The ability to convert the stored representation of images into a format usable for high-resolution display appears to be an essential component of most biomedical image applications. Additionally, the communications technology required for distributing images online will differ substantially, depending on functional requirements (for example, static vs. moving pictures). Therefore, the requirements associated with these forms of communications must be determined. Among the applications of image storage and transmission in biomedicine are:

- Full-text retrieval of journals by page, including images.
- Nontextual indexes for publications such as a chromosome map as an index to a genetics text.
- Preservation of texts by digital techniques.
- Picture archives. Test cases are being developed by the Library in the areas of rheumatology and dermatology.
- Digitization, storage, retrieval, and interpretation of radiographic and ultrasonic images.
- Diagnostic scans: radionucliotide, computerized tomography, positron emission tomography, nuclear magnetic resonance.
- Physiologic diagnostic material, such as EKG and pulmonary function graphs.
- Anatomical images e.g., photographs.
- Microscopic images, including both light and electron micrographs.
- Autoradiograms.
- Cell culture system analyses based on digitized images.
- Stereochemical three-dimensional images of macromolecules
- Multidimensional graphic analysis and depiction of modeling data.

Networking and Gateway Systems

Certain features of current computer networks raise issues for the Library in providing access to data bases outside its own computer system.

Computers and communications are coming together both technically and organizationally. The first networking experiments in 1965 highlighted needs leading to the development of ARPANET (Advanced Research Projects Area Network) by the Department of Defense. ARPANET was the first of the large packet-switched (rather than circuit-switched) networks permitting loadsharing between computers, message switching, and sharing of data and programs, as well as providing a variety of remote services. Recent estimates indicate that ARPANET users spend about one third less time on computing than would be the case if they did not have access to the shared resources of the network. Not surprisingly, user satisfaction with ARPANET is generally high.

In 1969, a technical landmark was reached when the computer costs associated with packet-switched networks fell below the expense of the communications involved. Since then, networking costs have continued to fall. Experience over the past 20 years has demonstrated that networks are feasible, cost-effective, and very reliable.

Networks permit the interconnection of independent devices that may be dissimilar and incompatible. Two kinds of networks are in widespread use:

- LANs (Local Area Networks) tend to cover a limited geographic area.
- VANs (Value Added Networks) generally cover the entire country, allowing cost-effective access to time-sharing systems, data bases, and information utilities. All current Library data bases are accessed through VANs (TYMNET, TELENET, UNINET).

As they exist today, computer networks are usable communications utilities, although marked by a number of drawbacks:

- Incompatibilities between networks impose a burden of accommodation on users.
- Using network-based utilities such as electronic mail requires training, since no consistent set of commands or functions is in use.
- In terms of biomedical information requirements, publicly available networking technology is adequate for text, but not for graphics.

However, work is in progress to overcome current networking limitations, and communications between computer systems, for either textual or nontextual signals, will not be impeded because of technical limitations in the future.

Thus, it appears that the electronic communications pathways will continue to improve whether or not the Library exercises any input to the process. The same cannot be said, however, of the higher level function of providing a consistent user interface for data base queries from dissimilar systems.

A functional description of a computerized information gateway has three components:

- Automatic connection to information resources residing in computer systems external to the Library's own computer, in an online, interactive mode.
- Automated source selection to one or more data bases to achieve the broadest basis for information retrieval.
- Automated mapping and reformatting of search queries to generate intellectually comparable searches in different systems.

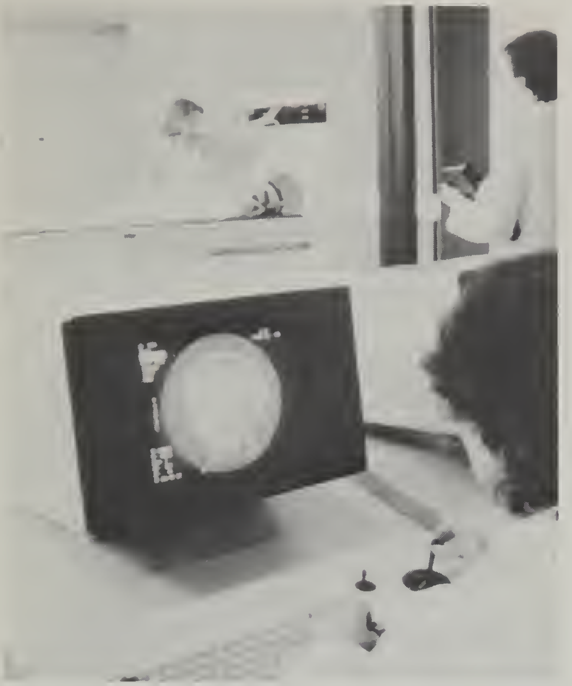
The Library's experience with CSIN and the current micro-CSIN project are good examples of freestanding gateway software that can take a user from one online system to another.

It is clear that the Library has an interest in gateway development. It also has an opportunity to exploit its own data bases to augment the state of the art.

Specialized Architectures

Specialized hardware architectures for very rapid search of textual data bases are beginning to emerge.[9] Hardware of this type that has come from the Library can typically locate complex patterns of several thousand characters in a large data base several times faster than even the fastest conventional computer hardware.[10]

Computer scientists are developing software for these machines that incorporates search strategies allowing some mismatches and other useful options. This hardware has the potential to permit the direct searches of data bases that now require extensive indexing. The result will enormously accelerate indexing of even the largest of current data bases.



Societal and Institutional Considerations

The Changing Information Age

Dramatic changes are occurring in the way individuals and institutions view information resources. Technology has been the principal force in making available and affordable improvements in information acquisition, processing, storage, retrieval, and delivery. The concerns of society about privacy, potential abuse, fraud, and security have posed the primary challenges to new information technologies.

A special challenge exists for the use of factual data bases in the biomedical and health fields: How to develop and use factual data bases to promote research, improve the health of all individuals, keep a lid on soaring medical costs, and not make the potential liability of factual data base providers too high for them to bear?

The challenge increases with the diffusion of small computers into homes and offices. This trend makes it possible to extend information services on demand to larger professional and nonprofessional segments of the population. In medicine, this development enables a computer-equipped lay public to query health data bases without the intervention of health professionals or library/information center intermediaries. Although it may take a decade or two to materialize, the probability is moderate to high that factual data base traffic will increase in this area.

Another concern with factual data bases deals with the content quality—the validity and reliability of the information contained in them. As early as 1963, the Presidential Science Advisory Committee strongly advised the development of critical data programs that offered users screened and useful information rather than long lists of titles and abstracts.[11] The Committee foresaw the value of specialized information centers staffed by subject specialists, aided by information experts who would provide quality data and information on request.

In subsequent years, much of the information community's energy was spent on controlling the explosion of scientific and technical literature and little on encouraging the growth of quality data bases. It would appear that the time has come to graft this philosophy of ''quality'' onto the reality of the Library's factual data base responsibility.

The security of information in factual data bases is also a matter of the highest importance. However, this problem is by no means unique or even special to the Library's interests. As noted elsewhere in this report, it is essential to protect the privacy of personal information in machine-readable patient records. However, privacy protection is significant to so many institutions and systems that the Library need not make a special effort to create specific techniques; they will come from other sources.

Indeed, an argument can be made that, for many of the factual data bases in which the Library has a legitimate interest, the principal thrust should be increasing content quality, accessibility, and making data base information widely available. That is the case currently for the Hazardous Substances Data Bank and for PDQ. It would also hold for data bases in basic science and clinical medicine where the public interest would best be served by easy access and wide availability.

Where questions of intellectual property rights enter, as with copyrighted and patented items, access may have to be controlled by fee or license considerations, but there still may be a role for the Library in improving availability. For example, in the realm of molecular biology or biotechnology, it is necessary to conduct literature searches before patenting a purportedly new plasmid. It would materially narrow the search if MEDLINE citations contained patent information. Correspondingly, the patent classification system and nomenclature bear little or no resemblance to biomedical terminology or taxonomy. Some sort of concordance or system of cross-mapping between the two domains would substantially increase availability of scientific information. A much more systematic review is warranted of the needs, opportunities, and impediments in this area.

Potential Liability

The potential liability of data base providers is based on tort principles and defined within the bounds of common law rather than statute. Specific torts that might be invoked in the case of harm resulting from information obtained through data bases fall into the following general categories:

- The fault of being careless.
- Strict liability, in cases where a product is defective and thereby dangerous.
- Misrepresentation, in cases where statements are made that information is accurate or complete and it proves otherwise.

To understand the extent of liability that might accrue to the Library, legal experts were consulted for a review of tort law and precedent. They made the following points:

- The Public Health Service mission is involved in life-sustaining science and information; thus, all deficiencies in the conduct of its mission may be viewed as potentially life-threatening.
- The Government's involvement in the area of health information imparts an aura of reliability to that information.
- The Public Health Service mandate for dissemination means that, if mistakes are made, they will be disseminated.
- Medical data bases will, by their nature, contain some information that is controversial and whose ultimate veracity is unresolved.
- International dissemination of that information will likely incur an international liability. If the United States is named as a defendant in an international tort suit, it must agree to be sued, and in that case, the action will be based on U.S. tort law.[12]

Furthermore, certain unique liabilities derive from the nature of computers and computerized information:

- Data base operators may create their own special liabilities, based on what it is possible to do with a computer system. For example, in a case against a van lines operator, the company was held to be negligent because its driver information and scheduling data base did not monitor whether drivers were exceeding Federal regulations prohibiting more than 70 hours of driving per week.
- Liability increases as a data base becomes the fastest available method to obtain information, particularly when the professional has no other source for the same information.
- Operators of private electronic bulletin boards have been held liable for improper or proprietary information (for example, credit card numbers) posted in their files.[13]

For these reasons, the Library can expect the data bases it develops and maintains to be held to the highest standard of care and the highest level of liability; that is, the standard will exceed those applied to medical data base efforts by commercial vendors.

The experts also addressed the individual liability of Library and NIH employees. It is noted that Public Health Service employees are excluded from personal liability for medical care they provide in the course of their assigned duties, including that provided in association with clinical trials. If an employee exercises a discretionary function, one that requires personal judgment, that person has immunity unless he violates some statute of which he should have been aware. Further, if an injured party sues the Government and loses, that party cannot then sue the individual involved.

The principal measures ordinarily used to protect against or minimize liability are:

- *Quality Control.* As noted above, there is an ongoing need to provide the best expert advice available by subject specialists throughout the life cycle of the data base.
- *Disclaimers.* Disclaimers alert the user that he must assume some liability; this concept is emphasized when the user is a highly educated professional. Although disclaimers do not eliminate liability, they should be used in any case to alert users to possible errors in the data.
- *Insurance.* This common method of managing liability risk in the private sector is unavailable to the Government because appropriated funds cannot be used for insurance. Hence, the Government is its own insurer.[14]

Finally, the users of factual data bases may also incur some liability. Once a data base is recognized as the most reliable or best source of information and is considered an essential part of the accepted standard of care, a practitioner may be judged negligent for failing to consult that source of information.

5 Observations and Recommendations

Observations

The National Library of Medicine is well-positioned and uniquely suited to provide guidance, standards, and support toward the development and distribution of biomedical factual data bases. Activities in this area are likely to become a dominant influence in gathering and communicating biomedical knowledge over the next 20 years. As the Library considers its distant goal, the following observations should guide its thinking:

- The Library's role as a provider of factual data base services will differ markedly from its responsibilities as a medical archive.
- The Library can best determine its potential role as a provider of factual data base services by entering into several experimental arrangements with other NIH institutes.
- To provide useful factual data base services, either domestically or internationally, the Library will need to pursue cooperative agreements with commercial enterprises.
- No realistic market analysis of demand is currently available for biomedical or public health factual data bases. Most ongoing activity is taking the form of Federal Government "push," instead of customer demand, or "pull." Federal budgets may soon be unable to support such proffered services.
- The Library's factual data base services will require an organizational infrastructure, a formal assignment of responsibilities, and a recognized budget and program.
- The Library will need considerable additional expertise and funding to support all the functions associated with factual data base services. Such functions include:
  - research and development, including data sources, full-text processing, and image processing (two-and three-dimensional);
  - current information/computer/communication support;
  - content and access controls;
  - maintenance services; and
  - legal services (validation, liability, malpractice).

Recommendations

Biomedical factual data bases and the Library's role in fostering their availability both have far-reaching potential. The technologies described can affect the decision-making and intellectual exploration capabilities of the health-care provider and researcher as profoundly as the automobile and airplane have extended our ability to move through our environment. Federal support of knowledge resources for biomedicine over the next two decades will, as much as or more than any national scientific endeavor currently under way or envisioned, lead to a rich harvest in our understanding and control of fundamental life processes.

The recommendations offered below highlight the areas of need and opportunity for the Library to attain this vision of the future. Although the recommendations are clustered within the principal areas of interest to the Panel, they are by no means exclusive to the context in which they are presented. Advances gained in one type of factual data base can and should be transferred to others, ultimately benefiting the technology as a whole.

A particular example of the need for cross-fertilization can be found in the specialized data bases required for biomedical research. There, the development of expert systems, networks and gateways, and specialized architectures enabling high-speed searches are all crucial to fulfilling the potential already apparent in this field. Therefore, in pursuing the actions recommended here, the Library should give strong emphasis to sharing lessons learned and goals achieved across the full spectrum of factual data bases.

Medical Practice-Linked Data Bases

(1) The Library should establish an intramural program for developing practice-linked data bases. The program should, among other things, promote factual data base standards, such as the Unified Medical Language System. Because the program will place additional responsibilities on the Library without diminishing its traditional mandate, funding should be sought from new appropriations rather than reprogramming existing resources. Wherever feasible, program costs should be shared with organizations responsible for data base content.

(2) Having established the program, the Library should actively promote its services within NIH and to the academic community in this country and abroad. The Library should develop models for sharing development costs and for ongoing cost reimbursement through licensing agreements with public agencies and private vendors.

(3) The Library should immediately and aggressively support the addition of automated information retrieval to the health professions curricula. This effort may be most useful over the next 5 to 10 years. After that, the increasing power of computer-based information systems, combined with growing computer literacy among health professionals, may obviate the need for direct Library support.

Patient Record Data Bases

(4) The Library should not attempt to play a substantial role in the development of ''health card'' technology or its applications. Further, the Library's role in corporate record-keeping developments should be limited to integration of the Unified Medical Language System.

(5) The Library should work with scientific and professional societies and other Government agencies to develop standards for clinical records contained in specialized data bases defined by disease entities or other research criteria.

(6) The Library should encourage the scientific exploitation of large clinical data files by providing extramural support for worthy investigations. Such grants should be evaluated on their own merits within the existing grant application process.

(7) The Library should indicate its willingness to store and make available appendiceal data files of selected published research, subject to conformity with standards developed in collaboration with journals and publishers.

(8) The Library should develop networks for providing online communication services to clinical practitioners. A networked health-care community is an integral part of the distant goal, but likely to become a standard of practice only in the second decade of the coming 20-year period. Medical librarians represent such a constituency today, however, and a pilot project involving them should begin immediately.

Biomedical Research-Oriented Data Bases

(9) The Library should immediately establish a program of biotechnology information to serve as both a repository and distribution center for this growing body of knowledge and as a laboratory for developing new information analysis and communications tools essential to the continued advancement of this field. As a part of this function, the Library should also immediately establish a liaison with biomedical research scientists involved in the creation or management of factual data bases. Special emphasis should be given to developing interlinkages, common retrieval strategies, and more capable analysis tools.

This should remain a high priority for the coming two decades. The gains to be realized in biomedical research, especially in the fundamental processes that control living cells, are likely to produce swift and pervasive advances in the biological sciences and the practice of medicine. In addition to molecular biology, other important trends in biomedical research, if accompanied by computerized information resources widely used by scientists, should be identified and similarly linked to the Library's literature retrieval services.

Taking advantage of its preeminent position in indexing the world's biomedical literature, the Library should sponsor meetings of scientists responsible for designing and maintaining research-oriented genetic factual data bases. The primary purpose for such gatherings would be to obtain consensus regarding the addition of terms to MeSH. A probable additional effect would be movement toward standardized indexing and retrieval methods.

(10) In instances where established, valuable biomedical research-oriented data bases might be lost for lack of continuing support, the Library should extend technical and financial assistance. The Library might also consider becoming the repository for endangered data bases.

Expert and Modeling Systems

(11) The Library should continue to support, through grants and other mechanisms, the development of expert and modeling systems in medical research, practice, and education. Development should be supplemented with field trials large enough to prove the utility of such systems in bettering the Nation's standard of health care.

(12) The Library should support research into the development and maintenance of computational models, both quantitative and qualitative, that interact with relevant factual data bases. In addition, the Library should begin to develop in-house expertise in modeling and simulation technology.

(13) A natural extension of the Library's role as a repository of printed textbooks for standard medical care will be the storage in electronic format of knowledge bases that include both data and medical lore. The Library should promulgate standards for those knowledge bases by investigating the current technology and choosing one or two candidates for archival-quality storage.

(14) The Library should develop specialized pseudo-English or menu-driven interfaces for certain factual data bases. Initially, one medical and one scientific data base should be chosen. The medical data base interface may, in fact, be subsumed by work on the Unified Medical Language System and its development costs viewed as an integral part of that effort.

(15) The Library should selectively support extramural research into full natural language systems for factual data bases, having first identified the need for such systems by medical researchers. The Library should explore the possibility of obtaining joint funding with other Government agencies that support similar research.

(16) The Library should begin an expert-system development project to model the behavior of an excellent medical reference librarian. It is strongly recommended that the Library use existing expert-system development hardware and software.

(17) The Library should carefully consider funding initiatives into data base introspection systems, realizing that this is state-of-the-art artificial intelligence research and unlikely to produce working systems for at least 5 to 10 years.

Technologies in Support of Factual Data Bases

(18) Working with representatives from medicine, biology, and communications, as well as the National Bureau of Standards and professional societies of the computer graphics industry, the Library should begin defining technical standards for biomedical image storage and transmission. Once defined, the standards can be compared with the image technology being developed by industry to determine if a unique window of opportunity (or obligation) exists for the Library in this area. In any event, historical performance suggests that if the Library does not develop such standards for the medical community, a proliferation of incompatible systems will result.

(19) The Library should support research into the application of artificial intelligence as a means to access several dissimilar factual data bases with a single query. Such independent-system searching will require analysis of hardware and software networking technology, identification of so-called intelligent agents (that is, automated systems smart enough to make the right choices) for source selection, and the development of a Unified Medical Language System.

(20) The Library should study and consider purchasing state-of-the-art data base search hardware for in-house experimentation and sharing with other NIH-supported resources. In addition, external research should be undertaken in the medical and scientific use of such equipment

References

1. Goldberg RN, Weiss SM. An experimental transformation of a large expert knowledge base. *J Med Syst* 1982;*6*:141-52.

2. Bernstein LM, Siegel ER, Koff RS, Merritt AD, Goldstein CM. The hepatitis knowledge base. *Ann Int Med* 1980;93:183-222.

3. McKusick VA. *Mendelian inheritance in man: Catalogs of autosomal dominant, autosomal recessive, and x-linked phenotypes.* 6th ed. Baltimore: Johns Hopkins University Press, 1983.

4. Kissman HM. Information retrieval in toxicology. *Ann Rev Pharmacol Toxicol* 1980;20:285-305.

5. Masys DR, Hubbard SM. Technical information programs of the National Cancer Institute. *Am Soc Info Sci* 1987; in press.

6. Gibson M. Major smart card products and installations. In: Levy AH, Williams BJ, eds. *Proceedings AAMSI 86—5th annual joint national congress.* Washington, DC: American Association for Medical Systems and Informatics, 1986:268-70.

7. Blum RL. Medical information science: its scientific and engineering aspects. *Med Inf (Lond)* 1984;9:214.

8. Kingsland LC III. The evaluation of medical expert systems: experience with the AI/RHEUM knowledge-based consultant system in rheumatology. In: *Proceedings of the Ninth Annual Symposium on Computer Applications in Medical Care.* Washington, DC: IEEE Computer Society Press, 1985;292-5.

9. Gabriel RP. Massively parallel computers: the connection machine and NON-VON. *Science* 1986;31:975-8.

10. Yu KI, Hsu SP, Otsubo P. The fast data finder—an architecture for very high speed data search and dissemination. In: *Proceedings of the Computer Data Engineering Conference.* (COMPDEC) April, 1984.

11. Weinberg A, et al. *Science, government, and information: the responsibilities of the technical community and the government in the transfer of information: a report of the President's science advisory committee*, Washington DC: U.S. Government Printing Office, 1963.

12. Paul Derensis, Chairman of the American Bar Association's Section on Tort Liability for Use of Computer Systems. In comments to the National Library of Medicine, long-range planning panel on obtaining factual information from data bases, November 18, 1985.

13. Robert Poling, Specialist in American Public Law, Congressional Research Service. In comments to the National Library of Medicine, long-range planning panel on obtaining factual information from data bases, November 18, 1985.

14. Robert Lanman, Office of the the General Counsel, U.S. Department of Health and Human Services. In comments to the National Library of Medicine, long-range planning panel on obtaining factual information from data bases, November 18, 1985.

Appendix A
Factual Data Bases in Basic Research

Overview

During the past few years, a number of online factual data bases that cover the the biological sciences have appeared. Many originated as manually maintained textual data bases that were transferred to machine-readable form with the advent of word processing. Some remain in word processing systems, while others have moved into the environment of a data base manager, usually taking advantage of relational data base software. GenBank is one well-known data base of great value to the molecular biology community.

GenBank attempts to maintain a collection of all DNA and RNA sequences that have been published in the scientific literature. A typical entry contains a series of fields describing the biological source of the sequence, the literature reference it was obtained from, the sequence itself, and a formal description of its biological interest and relevance. Most of the information comes from the original publication that described the sequence. Currently, a good deal of this annotation is presented in a standard, well-defined format that allows the information to be retrieved from the data base by a semi-intelligent computer program. However, such programs must be provided by GenBank's end-users; they are not part of the data base.

Uses and Users

Almost everyone in the field of molecular biology now deals with DNA sequence information. Just as chemists are concerned with the structure of their compounds, so molecular biologists are concerned with the structure, and hence the DNA sequence, of the genes that they study.

Of a special interest to most molecular biologists is the comparison of the sequence they are studying and all other known sequences. In particular, they are looking for homologies between sequences that might suggest functional similarities. For example, a molecular biologist studying a particular virus would almost certainly want to obtain the sequence of that virus. In the process, a number of genes would be identified, some of known function, others not. Immediately, the investigator would want to know if the sequence of a gene of unknown function resembled the sequence of another gene in GenBank whose function was known. Finding such a homology would suggest the two genes shared similar functions. The possibility of similar function could then be tested experimentally.

From an empirical standpoint, then, the most important piece of information in GenBank is the sequence of the gene. The second most important is the literature reference to the sequence. With that in hand, the investigator can go to the publication and find all the pertinent biological information about the sequence. Having a synopsis of the biological information available in GenBank saves some time, but is probably not a satisfactory alternative to consulting the original literature itself.

The application described above is probably the most widespread use of GenBank at this time, but it is by no means the only use. Other applications include the creation of sub-data bases containing particular categories of sequence, usually chosen from the annotation in GenBank.

Data Base Maintenance

As with most factual data bases available today, the maintenance of GenBank's content requires a team of specialists who scan the scientific literature to find publications with relevant information. For GenBank, that means identifying all publications containing descriptions of new DNA sequences. The GenBank team then obtains copies of those papers and extracts the pertinent information. Much of the data needed for GenBank are well defined and require little or no special training for identification.

An important component of GenBank, some may argue the most important component, is not so easily derived. Generally, the source publication does not present the annotation describing the information content of each sequence in a standard manner. In fact, the presentation is often so cryptic that not just one, but several specialists may be required to unravel it. This is the most time-consuming step in maintaining the data base.

This is also the point at which it becomes exceedingly difficult to keep the data base up to date. When a sequence is published, usually only a limited set of knowledge about that sequence is available. As time passes, further knowledge is gained, but may not be presented in the scientific literature as part of a sequence paper. Consequently, it does not come to the attention of the annotators. Incorporating this additional information into the data base then becomes a major problem, both strategically and philosophically: Who should do it?

Right now, it seems clear that manual retrieval and assembly of facts from the printed literature will prove impractical in the long run. Fortunately, several anticipated developments would eliminate the need for much of that. For example, most investigators already have their sequences in machine-readable form, either on a university mainframe or a small personal computer. As electronic networks become more accessible, it will be easy for those sequences to be transferred directly into GenBank. It may also be possible for GenBank to solicit information directly from the investigator. In that way, the data would be entered by the expert, with the managers at GenBank merely overseeing the effort.

Another scenario could unfold when the literature itself becomes available in machine-readable form. By then, it should be possible to develop semi-intelligent computer programs able to scan the literature and retrieve many of the facts required for GenBank. However, as suggested above, it seems unlikely that this will provide complete maintenance of the data base. Inevitably, some facts will still be buried within the literature or may never be published at all, residing solely in the investigator's notebook.

We might anticipate a time when facts within data bases constitute a form of publication acceptable to the basic research community. We are already seeing some indication of that in the DNA sequence field. When it does happen, the interaction between the data base managers and the individual investigator becomes the only means of incorporating the facts into the data base.

Still, updating the annotative aspects of the data base remains less straightforward. Certainly, the intelligent computer program of the future might be expected to retrieve annotative information from the literature. Existing programs can already identify much of the literature containing the additional facts. However, the subsequent retrieval of those facts is time consuming and inefficient when done by data base managers. Usually, it is the expert in a particular field who is best equipped to retrieve the basic facts quickly. Therefore, it seems likely that a major function of the data base manager of the future will be establishing and maintaining contact with the experts responsible for providing information to the data base.

Limitations

What are the limitations of the GenBank data base in its present implementation? The first is a certain unevenness in the depth to which sequences are annotated. Some entries contain many associated facts, while others contain only a few. This is a result of time pressures and staffing shortages. For the foreseeable future, the problem will be corrected only by providing additional funding for more staff, by soliciting unpaid help from experts in the user community, or by looking for more imaginative ways to collect and update the data.

A second problem is that, although many of the data have been entered in a form retrievable by appropriate software, a great deal of that information is not readily accessible. The major reason for this is that when any data base is designed, certain uses are anticipated and provided for in the programming. However, no existing data base management software is completely flexible. At the same time, restraints on flexibility lead to problems downstream, as do limitations in currently available hardware. Consequently, searches of the entire data base are time consuming, and the electronic links needed to establish relationships between facts in GenBank and those in other factual data bases simply do not exist.

The Future

One view of the GenBank of the future is that it will have two distinct categories of facts. One will be a collection of facts retrieved from the scientific literature by intelligent computer programs. The second will consist of facts supplied directly to GenBank by investigators. Those facts may never appear in the scientific literature, but they will nevertheless become part of it by virtue of their presence in GenBank.

Access to GenBank and other factual data bases will need to be rapid and easy. The data bases themselves will have to incorporate knowledge of related data bases, and inter-data base access must also be quick and easy. Perhaps one role for NLM, aside from merely cataloging the data bases available, would be to encourage and develop communication links between the libraries, data bases, and investigators.

Appendix B

Emerging Technologies: A View to the Future of Factual Data Bases

This appendix reviews the current state of technological development and the promise for the future in several areas of special relevance to the Library: data storage and transmission; machine understanding and processing of speech; image representation, storage, and transmission; and multimedia composite systems.

Digital Computer Storage Media Systems

The storage density of commonly available magnetic computer storage media systems may increase 5 to 10 times if the technical problems associated with vertical recording techniques (reading and writing multiple layers within a magnetic film) can be overcome. Magnetic systems will then have higher data storage capacities than comparable optical (''laser-disk'') storage. They also offer the benefit of being easily erasable, updatable, and reusable.

If vertical magnetic recording technology becomes sufficiently reliable within the next 5 to 10 years, the current trend toward the use of optical disks for high-capacity storage could be dramatically reversed. However, research in vertical magnetic recording technology is primarily in fixed disk design, whereas that in optical media technology involves removable disks. As a result, the two disk types will inherently fit different market niches and serve different storage purposes.

Widespread application of optical read-only memory disks for dissemination of data bases currently centers on the replicated CD-ROM (compact disk read-only-memory) in 120mm (4.72 inch) diameter disk sizes. CD-ROM is pressed from a master template, much as phonograph records are.

The CD-ROM format was originally developed for high-quality digital audio recording. Currently, the manufacturing capacity for CD-ROM is wholly absorbed by the entertainment industry. Industrial production is not expected to exceed that demand for at least one year. This has slowed the potential use of CD-ROM for data base publishing and has dampened its penetration of the data processing market. Further, the read-only character of replicated optical disks makes them most useful for applications that do not require updating capabilities.

The development of WORM (write-once, read-many-time) optical disk technology has been impeded by the lack of a suitable recording medium. As the optical disk currently used deteriorates with age, dependability decreases and the error rate increases.

Domestic manufacturers engaged in development of optical media technology include 3M, DuPont, and Kodak. Their expertise in continuous process industrial techniques may overcome the large-scale production problems that represent the major barrier now. It is expected that the devices available in the next 5 to 10 years will center around 8-, 5.25-, 3.5-, or 2-inch (approximate) disk diameters having data storage capacities of 40 to 10,000 megabytes. Proposed costs would run from $30-100 per disk, with disk drives ranging from $300 to $4,000.

Future prospects for optical digital data disk media and drive development include erasable technologies, such as magneto-optic, state change, and dye layer. However, they are likely to have 20 to 30 percent less data storage capacity than the write-once disks and greater expense than the read-only disks. Erasable products are expected to be commercially available in mid-1987, some of which will be compatible with the write-once and read-only disks. Most of the anticipated 3.5- and 2-inch media and drives will be erasable.

Audio

Most of the technical frontiers involving storage and transmission of audible signals are associated with the processing of speech. The development of speech compression techniques has resulted in steady but slow progress toward increased comprehensibility with reduced digital storage requirements. Current systems cost between $100 and about $10,000 and provide two-to-eightfold compression of speech with varying degrees of fidelity and speaker recognition. Voice mail systems are available commercially for all levels of computer capability, including microcomputers, but are not widely used.

Voice recognition programs of varying sophistication, with vocabularies of 10 to 1,000 words are commercially available or in laboratory use. However, they largely serve as control inputs for programs with a limited number of commands. Sophisticated applications such as very high-accuracy voice actuated typewriters await the development of advanced speech understanding capabilities, as described below, and appear to be at least 5 to 10 years away.

Substantial work in this area was done by DARPA in the 1970's. The program ended in 1976 because existing computer hardware was not powerful enough to produce a satisfactory response time; for example, a single sentence could take many minutes or even many hours to decode. In any event, the required hardware was too expensive.

DARPA's Strategic Computing Program is reopening that research because of decreased hardware costs due to integrated circuit technology. In addition, the prospect of obtaining realtime response using parallel processing techniques, which is a prerequisite for progress in semantic interpretation, is good.

Also, much progress has been made in the area of knowledge representation and acquisition techniques. Active research programs are under way at IBM; AT&T; Texas Instruments; BBN (Bolt, Beranek, and Newman); and Carnegie Mellon University. Promising methods include those based on graphic analysis of speech waveforms, acoustic phonetics analysis, and statistical models of speech-based grammars.

Image Storage and Transmission

Current technology provides *Time* magazine with the tools to digitally decompose a high-quality full-page photo image into 48 megabits of information, compress it into 1 megabit, transmit it over satellite data links, and reconstitute it for four-color printing. The electro-optical devices required for this process cost several hundred thousand dollars.

Currently available bit-mapped graphics displays, such as that in the Sun UNIX-based workstation, provide 1,000 x 1,000 bit resolution at a minimum cost of approximately $7,000 to $10,000. Too few economic incentives are in place at this time to promote the development of low-cost machines with higher resolution displays. However, CAD/CAM (Computer-Assisted Design/Computer-Assisted Manufacturing) users believe a good market exists for 2,000 x 2,000 bit resolution displays, and this use should drive prices down. Up to 8,000 x 8,000 bit resolution has also been identified, but the projected market for these super high-resolution displays will probably be limited to certain medical and other specialized applications.

The Patent Office is currently participating in a project that uses optical disk storage devices to deliver the equivalent of one full page of text and graphics to 50 workstations every two seconds. The success or failure of this project may influence the Government's willingness to fund continued technical development in the area of image transmission other than for defense purposes.

The Defense Department is supporting electronic publishing and distribution of information through its ''Computer-Aided Logistics Systems'' effort. Included are online graphics image systems to support the performance, maintenance, components, and use of weapons systems.

Multimedia Composite Systems

Digital storage and retrieval systems capable of integrated handling of text, images, and speech are in development. Within most of those systems, related information is processed as serial presentation of different kinds of signals: Terminals with the proper configuration can display formatted text with embedded graphics images and speech commentary. Terminals with lesser capability might display only text, so the system accommodates varying levels of hardware.

In some systems, the user may also record verbal comments and incorporate them at any point in the text. Nested systems of this design allow definition of any text, image, or audio as a hierarchical successor to any other component. Various manufacturers, such as Xerox and BBN, are major commercial developers of this technology.

Appendix C:
NLM Planning Process

In January, 1985 the Board of Regents of the National Library of Medicine resolved to develop a long range plan to guide the Library in wisely using its human, physical, and financial resources to fulfill its mission. The Board recognized the need for a well-formulated plan because of rapidly evolving information technology, continued growth in the literature of biomedicine, and the need to make informed choices of intermediate objectives that would lead NLM toward its strategic, long range goals. Not only would a good plan generate goals and checkpoints for management, actually a map of program directions, but it would also inform the various constituencies among the Library's users about the future it sought and could help to enlist their support in achieving that future.

At the Board's direction, a broadly based process was begun involving the participation of librarians, physicians, nurses, and other health professionals; biomedical scientists; computer scientists; and others whose interests are intertwined with the Library's. A total of 77 experts in various fields accepted invitations to serve on one of the five planning panels. Each panel addressed the future in one of the five domains that encompass NLM's current programs and activities. The domains, which provided the panels, a framework for thinking about the future are:

1. Building and organizing the Library's collection
2. Locating and gaining access to medical and scientific literature
3. Obtaining factual information from data bases
4. Medical informatics
5. Assisting health professions education through information technology

The Library chose a planning model with three components. First, it incorporates a general, somewhat indistinct vision of the future 20 years from now in medicine, library and information science, and computer-communications technology. That environment cannot be forecast precisely, but we can speak of a ''distant'' goal. That goal is seen as a societal objective whose attainment involves many organizations and agencies. NLM has a major role to play in achieving the goal and must plan its part. Second, while the 20-year goals are indistinct, there are opportunities for and impediments to achieving them. The opportunities and impediments can be more clearly envisioned because they appear to lie roughly 10 years away. Third, the specific steps that should be taken to remove the impediments and take advantage of the opportunities should be programmed for 3 to 5 years.

The planning process also involved participation within the Library. The Director provided his version of the future in the form of a ''Scenario: 2005,'' which was distributed to panel members and Library staff. NLM staff prepared background documents that reported NLM achievements in the five domains, and reviewed current planning. Senior NLM staff members also acted as resource persons to the planning panels.

At the end of the planning process, each panel formulated recommendations and priorities for future NLM programs and activities in the domain under its purview. The five panel reports were reviewed by the Board of Regents in June 1986. The Board then asked the NLM staff to analyze and reconcile their findings, eliminating any duplications and consolidating the recommendations. Together with the planning panel reports, this synthesized plan presents the official Long Range Plan of the Board of Regents of the National Library of Medicine.

Photographs were obtained from the several Bureaus, Institutes, and Divisions of the National Institutes of Health (including the Office of the Director, NIH, the Warren G. Magnuson Clinical Center, and the National Institute on Aging), the Uniformed Services University of the Health Sciences, the World Health Organization, and William A. Yasnoff, M.D., Ph. D..

December 1986